APPLICATION

FOR

UNITED STATES LETTERS PATENT

APPLICANT NAME: Bradford et al

TITLE: ACTIVE FLOW MANAGEMENT WITH HYSTERESIS

DOCKET NO. RPS920030131US1 (IRA-10-5791)

INTERNATIONAL BUSINESS MACHINES CORPORATION

CERTIFICATE OF MAILING UNDER 37 CFR 1.10

I hereby certify that, on the date shown below, this correspondence
is being deposited with the United States Postal Service in an
envelope addressed to the Mail Stop Patent Application, Commissioner for
Patents, P.O. Box 1450, Alexandria, VA 22313-1450 as "Express Mail Post
Office to Addressee"

Mailing No. _EV385165765US_____
on FEBRUARY 19, 2004
Karen Orzechowski
Name of person mailing paper
Signature: Karen Orzechowski
Date: 2/19/2004

ACTIVE FLOW MANAGEMENT WITH HYSTERESIS

FIELD OF THE INVENTION

This invention relates in general to bandwidth allocation and active flow management techniques for computer networks, and more particularly to implementation of hysteresis in active flow management techniques to increase network system throughput by making better use of available queue capacity.

BACKGROUND OF THE INVENTION

In computer network systems, active flow management techniques are commonly used to control the subscription and offered load of each thread data flow, along with the service rate of the network system itself, to achieve fairness of bandwidth allocation.

However, unnecessary packet drop due to short burst traffic may occur, and problems arise if there is no mechanism provided to preferentially treat short bursts of packets. Most active queue management algorithms drop some packets when congestion is detected, and indeed in an initial burst to detect incipient congestion. If the burst is sustained for a very short period, this can cause unnecessary packet drops because there is enough space in the packet buffer to be able to accommodate the burst. This is especially detrimental in Transmission Control Protocol (TCP) networks because each packet drop causes TCP retransmissions which can lead to very low useful throughput.

In TCP networks, it is known to use Explicit Congestion Notification (ECN) to mark packets and indicate to a sender that a congestion window should be adjusted to a lower rate. However, ECN applied in the case of very short and sustainable bursts can be detrimental to the total throughput because it unnecessarily causes the window to adjust when the packets in the burst could well have been transmitted only with a little price in latency. It is also known to use an exponentially weighted moving average of a queue level to smooth out bursts; however, this solution is computationally expensive.

What is needed is a method and system for active flow management for computer networks that sustains short burst packet traffic without causing unnecessary packet drops and at the same time not degrading the network system throughput for persistent bursts of packets, and which can be implemented in hardware without too much logic overhead.

SUMMARY OF THE INVENTION

The present invention provides for a computer network method and system that applies "hysteresis" to an active queue management algorithm. If a queue is at a level below a certain low threshold (L) and a burst of packets arrives at this network node, then the probability of dropping the initial packets in the burst is recalculated, but the packets are not dropped. However, if the queue level crosses beyond a "hysteresis threshold" (Ht), then packets are discarded pursuant to a drop probability. This allows more packets from the burst to get into the queue. Where the burst lasts for a short time (a "short burst"), then the present invention provides the ability to transmit every single packet.

According to the present invention, when a queue level is beyond the hysteresis threshold, and arrival rate into the queue is less than the sending rate from the queue, then queue level is decreased until it becomes less than the 'hysteresis threshold' (Ht). However, during this time, packets get dropped as per the drop probability until the queue level decreases to at least the low threshold (L). Thus, the present invention is intended to improve network performance where a burst is received into a queue when the queue level is low.

In one embodiment of the present invention, an adaptive algorithm is also provided to adjust the increment and decrement of transmit probability for each flow, together with hysteresis to increase the packet transmit rates by using packet data store to absorb bursty traffic. The proposed algorithm maintains the throughput for persistent bursts of packets. Using straight forward implementation of hysteresis in the active flow management will increase the system throughput by making better use of available queue capacity. This results in an increase in the queue level peak, potentially exposing the system to tail drops when subjected to severe bursts for bursty traffic. With the addition of adaptive increment and decrement of transmit probability of each flow, queue peak can be limited to a reasonable level, thus preventing tail drop.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 is a graphical representation of the relationship of transmit probability to buffer capacity according to the present invention.

Figure 2 is a flowchart illustrating a hysteresis algorithm according to the present invention.

Figure 3 is a graphical representation of results for queue occupancy for hysteresis factor levels according to the present invention.

Figure 4 is a graphical representation of the transmit rates for various levels of hysteresis according to the present invention.

Figure 5 is a table of transmit rates for bursty traffic according to the present invention.

Figure 6 is a graphical representation of a transmit rate throughput comparison for bursty traffic for RED, RED with hysteresis, SARED and SARED with hysteresis according to the present invention.

Figure 7 is a graphical representation of a latency comparison for bursty traffic for RED, RED with hysteresis, SARED and SARED with hysteresis according to the present invention.

Figure 8 is a high-level block diagram of a standalone network simulator appropriate for use with the present invention.

Figure 9 is a graphical representation of UDP traffic on the standalone network simulator of Figure 9 according to the present invention.

Figure 10 is a graphical representation of UDP short burst bursty traffic on the standalone network simulator of Figure 9 according to the present invention.

Figure 11 is a graphical representation of TCP constant traffic on the standalone network simulator of Figure 9 according to the present invention.

Figure 12 is a graphical representation of TCP bursty traffic on the standalone network simulator of Figure 9 according to the present invention.

Figure 13 provides tables of packet drop and packet transmit statistics according to the present invention.

Figure 14 is a flowchart that shows an adaptive algorithm according to the present invention.

Figure 15 is a graphical representation of aggregate transmit rates according to the present invention.

Figure 16 is a graphical representation of average latency according to the present invention.

Figure 17 is a graphical representation of aggregate transmit rates according to the present invention.

Figure 18 is another graphical representation of aggregate transmit rates according to the present invention.

Figure 19 is a graphical representation of aggregate transmit rates for steady traffic according to the present invention.

Figure 20 illustrates an embodiment of the invention tangibly embodied in a computer program residing on a computer-readable medium or carrier.

DESCRIPTION OF THE PREFERRED EMBODIMENTS

The present invention applies "hysteresis" to active queue management techniques. Hysteresis is generally defined as "the lagging of an effect behind its cause" and it is well-known to use hysteresis behavior in applications that switch between two transmission modes in a variety of network system applications. In the present invention, where a queue is at a level below a certain low threshold (L) and a burst of packets arrives at this network node, then the probability of dropping the initial packets in the burst is recalculated, but the packets are <u>not</u> dropped. However, if the queue level crosses beyond a "hysteresis threshold" (Ht), then packets are discarded pursuant to a drop probability. This allows more packets from the burst to get into the queue. Thus, where the burst lasts for a short time (a “short burst”), then the present invention provides the ability to transmit every single packet.

According to the present invention, when a queue level is beyond the hysteresis threshold, and arrival rate into the queue is less than the sending rate from the queue, then queue level is decreased until it becomes less than the 'hysteresis threshold' (Ht). If the arrival rate into the queue is less than the sending rate from the queue, and packets arriving into the queue are not discarded, the queue level will eventually rise and, after a while, no more packets will be accepted into the queue. According to algorithms taught by the present invention, arriving packets are randomly discarded but only to the point where the queue level reaches the hysteresis threshold. However, during this time packets get dropped as per the drop probability until the queue level decreases to at least the low threshold (L). Thus, the present invention is intended to improve network performance where a burst is received into a queue when the queue level is low.

Random Early Detection is a well-known queue management algorithm that is used in network routers to detect incipient congestion based on queue occupancy. As per the definition of RED, a low (L) and a high (H) threshold are defined for the queue. An exponentially weighted average (Qavg) of the queue occupancy is used to define congestion. If Qavg is less than L, the arriving packet is not discarded. If Qavg is greater than H, the arriving packet is discarded. If Qavg lies between L and H, a discard probability is calculated for the arriving packet based on the linearly increasing probability between L and H.

One of the drawbacks of RED is that it is not easy to set the thresholds to achieve optimum queue occupancy and network performance. Referring now to Figure 1, a graphical representation 10 of the relationship of transmit probability p to buffer capacity Q according to the present invention is illustrated, where T represents the transmit probability; T_0 represents an initial value for the transmit probability; H and L represent the High and low thresholds, respectively, for Random Early Detection (RED); and Qmax represents the buffer capacity. In the example illustrated, L = (1/4)*Qmax; H = (31/32)*Qmax, 0 <= L <= Ht <= Qmax; $T = 1 - (Q - L) *(1 - T_0) / (H - L)$, and $T_0 = 1/8$. However, the present invention is not limited to these values, and it is to be understood that other values may be selected for use with the present invention. The solid line 12 is the transmit probability curve when the hysteresis flag is turned OFF. The dotted line 14 is the transmit probability curve when the hysteresis flag is turned ON. If hysteresis flag (HT) is on, then discards should be delayed; otherwise, discard probability should be calculated as per the queue management scheme. Thus, according to the present invention, the queue level is treated as a vector instead of a scalar: i.e., the queue level now has a direction 16 along with a value 12 or 14.

Figure 2 is a flowchart illustrating a hysteresis algorithm 301 for setting the hysteresis flag according to the present invention. At step 302, the hysteresis flag (HT) setting is ascertained. If the HT is on, then the queue level is checked at step 304 to determine whether it has increased beyond or is equal to the hysteresis threshold (Ht). If yes, then the HT flag is switched OFF at step 306. Otherwise, the HT flag is kept ON at step 308.

Alternatively, if the HT is OFF at step 302, then the queue level is checked at step 310 to determine whether it has decreased to a value below the low threshold (L): if not, then the HT flag remains switched OFF at step 306; if so, then at step 312, it is determined whether the offered load (OL) is below the link capacity (C). Offered load is defined to be the aggregate traffic bandwidth being presented to the link between two nodes in a network by a data transmitter. Link Capacity is defined as the maximum bandwidth supported by a physical link between two nodes in a network. If OL is below C at step 312, then the HT flag is switched ON at step 308 ON (this implies that the incoming flows are well behaved); otherwise, the HT flag remains switched OFF at step 306.

The hysteresis algorithm 301 described above has been tested under constant traffic load as well as under bursty traffic load. Embodiments of the present invention have been implemented on two separate models: a Network processor simulation engine (NPSim) and an independent Network simulator (ns-2). As shown in Figures 3 et seq, there is considerable improvement in network throughput by using hysteresis on RED according to the present invention. Under the constant load conditions, traffic sent into the NPSim network processor simulation model running the algorithm 301 oversubscribed the link by 50%. A hysteresis factor (n) is defined as the ratio of the hysteresis threshold to the low threshold of RED, thus Ht =n*L. Results for queue occupancy for four levels of the hysteresis factor are shown in Figure 3: level 402 where n=1; level 404 where n=1.25; level 406 where n=1.5, and level 408 where n=2. Note that applying hysteresis to a constant load does not affect the throughput of the flow control scheme adversely. In fact, the initial hump 410 for case 408 where n=2 helps to slightly improve the throughput.

Figure 4 illustrates the transmit rates for various levels of Hysteresis. For case C1 502, the offered load (OL) is 50% of link capacity (C). For case C2 504, the offered load is 100% of link capacity. And for case C3 506, the offered load is 150% of link capacity. Note that for C3 506, 66.7% of traffic is transmitted.

Figure 5 illustrates transmit rates for bursty traffic; 150% of link capacity is sent during the BURST ON period, and 20% of link capacity is sent during the BURST OFF period. The BURST ON and OFF periods are varied for different levels of oversubscription.

Figure 6 illustrates a transmit rate throughput comparison for bursty traffic for RED 706, RED with hysteresis 704, Shock Absorber Random Early Detection (SARED) 708 and SARED with hysteresis 704. Further information about SARED is set forth in commonly-assigned U.S. patent application entitled "FLOW CONTROL IN COMPUTER NETWORKS", serial number 10/160,507, filed June 3, 2002, which is incorporated herein by this reference. For Both RED and SARED queue management schemes, there is a considerable improvement in the transmit rates for bursty traffic when hysteresis is applied to the scheme.

Figure 7 illustrates a latency comparison for bursty traffic for RED 804, RED with hysteresis 802, SARED 810 and SARED with hysteresis 806. This illustration shows the price that needs to be paid in terms of additional queue occupancy (which can be translated into latency) for achieving the higher throughput. In all cases, the queue occupancy increases only marginally.

Figure 8 is a block diagram of a standalone network simulator NS-2 902 for use with the present invention. A node *s1* 904 sends traffic to a node *s2* 906 via router *r1* 910. An active queue management algorithm (RED) runs at *r1*, with a packet size of 1000 bytes. Note that the *r1*-*s2* link 912 is oversubscribed by 50%, with a buffer size at *r1* of 600 packets, and Ht = L for no hysteresis, Ht = 2*L* for hysteresis.

Results of testing the hysteresis algorithm 301 on a User Datagram Protocol (UDP) traffic first Case 1002 on the NS-2 902 are illustrated in Figure 9. There are 2 bursts 1004 and 1006 in the simulation. Each burst 1004 and 1006 lasts for 2 seconds. An OFF period 1008 between the burst 1004 and 1006 is also two seconds, resulting in queue occupancy for RED 1010 and RED-with-hysteresis 1012 as shown.

Results of testing the hysteresis algorithm 301 on a UDP short burst bursty traffic Case 1102 on the NS-2 902 are illustrated in Figure 10. Two bursts 1104 and 1106 are run in the simulation. The first burst 1104 lasts from 0 sec. to 0.5 sec. and the second burst 1106 lasts from 3.0 sec. to 4.0 sec., with resultant queue occupancy for RED 1108 and RED-with-hysteresis 1110.

Figure 11 illustrates TCP Tahoe constant traffic on the standalone network simulator 902, wherein queue occupancy is shown for RED 1202 and RED-with-hysteresis 1204.

Figure 12 illustrates TCP Reno bursty traffic on the standalone network simulator 902. Two bursts 1302 and 1304 each last for one second, with an intervening OFF period

1306 of one second, wherein queue occupancy is shown for RED 1308 and RED-with-hysteresis 1310.

Lastly, Figure 13 provides tables of packet drop and packet transmit statistics according to the present invention.

In another embodiment of the present invention, an adaptive algorithm is also provided to adjust the increment and decrement of transmit probability for each flow, together with hysteresis to increase the packet transmit rates by using packet data store to absorb bursty traffic. The proposed algorithm maintains the throughput for persistent bursts of packets. Using straight forward implementation of hysteresis alone in the active flow management will increase the system throughput by making better use of available queue capacity. However, this results in an increase in the queue level peak, potentially exposing the system to tail drops when subjected to severe bursts for bursty traffic. With the addition of adaptive increment and decrement of transmit probability of each flow, queue peak can be limited to a reasonable level, thus preventing tail drop.

An embodiment of the invention thus proposes an algorithm including the following two components to improve the performance of active flow management: (1) a Bandwidth Allocation Transmit (BAT) algorithm, without SARED but with hysteresis; and (2) an adaptive transmit fraction Ti responsive to certain conditions (e.g., queue and/or traffic).

Further information about BAT is set forth in commonly-assigned U.S. patent applications entitled "METHOD AND SYSTEM FOR PROVIDING DIFFERENTIATED SERVICES IN COMPUTER NETWORKS", serial number

09/448,197, filed Nov. 23, 1999; and "METHOD AND SYSTEM FOR CONTROLLING FLOWS IN SUB-PIPES OF COMPUTER NETWORKS", serial number 09/540,428, filed March 31, 2000, both of which are incorporated herein by this reference.

With regard to (1) the first part of the two-part algorithm (BAT without SARED but with hysteresis), the Transmit fraction of BAT for flow i, Ti, is defined as follows:

```
If $fi(t) <= fi,min$            then $Ti(t + dt) = min(1, Ti(t) + w)$;
else if $fi(t) > fi,max$        then $Ti(t + dt) = Ti(t)(1-w)$;
else if $B(t) = 1$              then $Ti(t + dt) = min(1, Ti(t)+CiBavg(t))$;
otherwise                       then $Ti(t + dt) = Ti(t)(1-DiOi(t))$;
```

where Ci and Di are constants used for increment and decrement, respectively, of Ti. Ci and Di are defined by subscription of each flow, fi,min, and the service rate of the system, S. They are given as follows:

$Ci = (S + fi,min-(fl,min + f2,min +... + fn,min))/16$; and

$Di = (S - fi,min)*4$.

Hysteresis is incorporated according to the following algorithm: if hysteresis is on and the queue level is less than the hysteresis threshold, then no packet will be dropped -- i.e., Ti is updated but does not apply to packets; else, if hysteresis is off, then packets are processed as normal -- i.e. Ti is applied to each packet.

With regard to (2) the second part of the two-part algorithm (adaptive transmit fraction Ti based on certain conditions), prior art implementations of BAT have been guarded by SARED which will reduce Ti when queue occupancy exceeds the SARED threshold, e.g., 25% of maximum queue capacity. With hysteresis, there is no need for

SARED to guard BAT. However, this may increase the queue level peak which may cause tail drop due to high queue occupancy. In order to prevent packets from tail drop, the present invention provides for an adaptive increment and decrement of transmit probability of each flow, Ti, to prevent tail drop while maintaining the advantage of hysteresis, e.g. higher transmit rates with bursty traffic. An embodiment of the present invention comprises a normal Ti algorithm with an extended Ti algorithm to adapt Ti for good conditions (low queue and/or light traffic) and severe conditions (high queue and/or severe traffic), respectively.

For conditions between good and severe, an adaptive increment and decrement of Ti is used based on the condition of traffic or the direction the queue level is moving. "Severe conditions" implies that no amount of congestion control will be able to prevent the discard of arriving packets. Referring again to Figure 1, this is the case where the queue level (on the x axis) has reached the high threshold (H), beyond which the drop probability is = 1. Between L and H (i.e. between the good and severe conditions) is where the flow control lies – and with regard to (2), this is an adaptive increment and decrement of Ti unlike a fixed change in Ti as was done in (1). The advantages of the proposed algorithm include using queue data store to absorb short burst traffic to achieve higher throughput, and adjusting increment and decrement of transmit probability for each flow to limit queue peak to prevent tail drop.

Figure 14 illustrates an adaptive algorithm 1502 according to the present invention of Ti using hysteresis as good or severe condition indicator for flow i, given constants Ci and Di calculated from subscription of flow i and service rate of the system.

In step 1504, the HT flag is determined. If HT is ON, then in step 1506 Ti is computed using Ci and Di in BAT. Else, if HT is OFF, then in step 1508 Ti is computed using F(Ci) and G(Di) in BAT, where F is a decreasing function and G is an increasing function. One embodiment sets F(C) = C/2 and G(D) = min(1, 2*D); however, other embodiments may utilize different values, and the present invention is not limited to these values. After step 1506 or step 1508, in step 1510, the Ti is updated.

HT is set according to algorithm 301 of Figure 2. Given low threshold L, high threshold H, and hysteresis threshold Ht where 0<=L <= Ht <= H <= maximum queue capacity. Hysteresis is ON initially. It is preferred that both Ti and HT are updated periodically and the period is dependent on queue size and service rate. When hysteresis is ON and the queue level is less than the hysteresis threshold, no packet will be dropped. When hysteresis is OFF, Ti is applied to every packet.

The benefit of the present invention is demonstrated by simulation results shown in Figures 15 through 19. These simulations use hysteresis as an indicator of good and severe conditions. In these simulations, F(Ci) = Ci/(2^3) and G(Di) = min(l, 2^3*Di) are used. The hysteresis threshold Ht is 2*(low threshold). The simulation contains seven cases and each has constant burst OFF duration for 1.07 ms except for the case of 100% burst ON duration. The respective "burst ON" durations for the seven cases are: case 1602 - 100%; case 1604 - 66.6%; case 1606 - 61.5%,;case 1608 - 54.5%; case 1610 - 50.0%; case 1612 - 44.4%; and case 1614 - 37.4%. For example, the traffic pattern for the 66.6% burst ON duration case 1604 is made by the cycles of 2.13 ms burst ON and followed by 1.07 ms burst OFF. The traffic pattern is 150% subscription when bursty traffic is ON and 20% subscription when bursty traffic is OFF.

The simulations shown in Figures 15 through 19 contain four (4) UDP flows, and the subscriptions of each flow are: 10%, 10%, 20%, and 0% of total bandwidth, respectively. The offered load for each flow is 50%, 40%, 20%, and 40% of total bandwidth, respectively, when burst is ON and 5% for each when burst is OFF. Note that the fourth flow is best effort – it will only avail of the bandwidth that is left over after allocation to the remaining flows. According to a min-max algorithm, the ideal transmit rates for each flow under constant case (100% burst ON) are 60%, 75%, 100%, and 50%, respectively.

The aggregate transmit rates are illustrated in Figure 15, and average latency for the same cases are shown is in Figure 16 for various bursty traffic. Figures 15 and 16 show that the proposed invention can result in higher transmit rates for bursty traffic while having higher latency because of the use of queue capacity to absorb bursty traffic.

Specific transmit rates for individual flows obtained through the algorithms of the present invention 1802 and through prior art BAT methods 1804 are illustrated in Figures 17, 18, and 19. Figure 17 is the aggregate transmit rate for each flow for 50.0% burst, and Figure 18 is the aggregate transmit rate for each flow for 66.7% burst.

Figures 17 and 18 show higher transmit rates for each flow for different burstyness. They illustrate that the present invention obtains higher transmit rates for bursty traffic through better use of available queue capacity to accommodate bursty traffic. Figure 19 is the aggregate transmit rate for each flow for steady traffic (100%

burst). It shows that the algorithm according to the present invention can maintain the same transmit rate for steady, or persistent bursty, traffic.

Overall, the present invention can achieve higher aggregate transmit rates for a variety of traffic burst characteristics by making better use of queue capacity and can maintain the level of performance for persistent traffic.

The algorithm for hysteresis with adaptive increment and decrement of transmit rate can also be easily applied to Weighted RED (WRED) to achieve higher transmit rates. In this case, a different Ht threshold could be defined for each of the flows subscribing to the available bandwidth along with the individual definitions of their low (Li) and high (Hi) thresholds. When the hysteresis flag is turned ON, the probability of dropping can be decreased by twice of what it would be when the hysteresis flag is turned OFF, thereby accepting more packets into the queue when there is less congestion.

The invention may be tangibly embodied in a computer program residing on a computer-readable medium or carrier, such as the floppy disc 2105 or hard drive 2101 shown in Figure 20. The medium 2105 may comprise one or more of a fixed and/or removable data storage device, such as a floppy disk or a CD-ROM, or it may consist of some other type of data storage or data communications device. The computer program may be loaded into the memory 2102 of a network manager computer device 2110 for execution. The computer device 2110 may be connected to a network via network interface 2103. The computer program comprises instructions which, when read and executed by the computer device 2110, causes the computer device 2110 to perform the steps necessary to execute the steps or elements of the present invention.

The foregoing description of the exemplary embodiment of the invention has been presented for the purposes of illustration and description. It is not intended to be exhaustive or to limit the invention to the precise forms disclosed. Many modifications and variations are possible in light of the above teaching. It is intended that the scope of the invention be limited not with this detailed description, but rather by the claims appended hereto.